CSC110 Project Written Report: r/HowCovidImpactsMentalHealth

Asma Jaseem, Chloe Lam, Jin Yi (Helen) Li, Sarah Yi Xu

Tuesday, December 14, 2021

Problem Description and Research Question

Since the eve of 2020, the global spread of the coronavirus disease, or COVID-19, drastically altered daily life for people around the world, introducing new restrictions and guidelines for everyone to follow in an attempt to prioritize public health. Many governments around the world looked to lockdown as their first solution to limiting the spread of COVID-19 and thus introduced public health policies to restrict the amount of social gatherings (Government of Canada, 2021). As people remained disconnected for long periods of time, concerning problems like isolation, loneliness, anxiety, and even depression became increasingly more prevalent amongst the general population. A study published in *The Lancet*, a medical journal, estimates that the pandemic has caused a 28% global increase in depression compared to pre-pandemic, while anxiety had a 26% rise globally (COVID-19 Mental Disorders Collaborators, 2021). Conducted in July 2020, a Kaiser Family Foundation Poll reported further that due to worrying and stressing over COVID-19, 36% of the participants have struggled with sleeping or eating, 12% have increased their substance use or alcohol consumption, and another 12% have experienced worsening chronic conditions (Kamal et al., 2021). At the same time, however, mental health care became more limited, making it harder for people to seek relevant professional support. According to the World Health Organization, the pandemic has disrupted mental health services in 93% of the world, with 67% of these countries experiencing disruptions in counseling and psychotherapy (World Health Organization, 2020).

As the public focuses primarily on the details of any new confirmed cases or deaths, the social impacts of COVID-19 have been overshadowed by these numerical facts. Specifically, the pandemic's effects on individual mental health have received relatively less public attention. However, as everyone faces a variety of pressure and stress on a daily basis, remaining mentally healthy is key to living a happier and more balanced life. As people are forced to spend more time at home to keep themselves and others safe, individual mobile and internet usage spiked along with significant increases in social media exposure (Andrews, 2020). Taking advantage of this increase in online presence, this project aims to answer the question **'How does social media reveal the impacts of COVID-19 on mental health?' This project defines mental health through the terms of depression, anxiety, ocd, obsessive compulsive order, panic attack, insomnia, mental health, counseling, and psychiatry.** By analyzing relevant data taken from online platforms, like Google and Reddit, this project examines the way public attention toward mental health and the mental wellbeing of online communities have changed as a result of the pandemic. Specifically, sentiment analysis will be performed to intelligently analyze the emotional state and attitude of Reddit users before and during the coronavirus outbreak (Mohammad, 2016). Two common approaches in sentiment analysis are polarity-based, which classifies a piece of text as positive, negative, or neutral, and valence-based, which calculates the intensity of the sentiment (Mohammad, 2016). For instance, a polarity-based approach would classify both "happy" and "very happy" as positive while a valence-based approach would yield a greater positive intensity value for "very happy" than for "happy." Both approaches will be utilized for this project.

Dataset Description

There are 3 relevant datasets.

1. Daily global confirmed cases

   In csv format, this dataset comes from John Hopkins University's COVID-19 dashboard, which uses data from the World Health Organization (Dong et al., 2020). It includes the number of confirmed cases worldwide from January 22 to December 30, 2020. In the file, rows represent countries while columns represent dates, so each cell reflects the total number of cases up to that date in that country.

   Sample data (first 7 columns):

Province/State	**Country/Region**	**Lat**	**Long**	**1/22/20**	**1/23/20**	**1/24/20**
Victoria	Australia	-37.8136	144.9631	0	0	0

2. Search interest of mental health-related terms

   In xlsx format, the second dataset is collected by a Kaggle user using Google Trends and contains the search interest of 9 mental health related terms (Hao, 2020). Stored in separate files, data from June 2015 to May 2020 is included for Canada, and data from June 2019 to May 2020 is included for US, Italy, Iran, Japan, South Korea, UK, and around the world. Search interest is represented by either a number from 1 through 100, representing the least popular to the most, or just 0 if there is insufficient data.

   Sample data for global search interest (first row):

Week	**depression**	**anxiety**	**obsessive compulsive disorder**	**ocd**	**insomnia**	**panic attack**	**mental health**	**counseling**	**psychiatrist**
2019-06-16	81	87	61	58	75	72	45	99	86

   Note: Data from Japan and South Korea are identical. Assuming it is a data collection error and since the specific country does not affect our program, only one of them (Japan, randomly chosen) is kept.

3. Posts from mental health-related subreddits

   In csv format, the third dataset includes posts scraped from 8 mental health related subreddits—r/depression, r/Anxiety, r/OCD, r/insomnia, r/PanicAttack, r/mentalhealth, r/counselling, r/Psychiatry—using pandas and pmaw libraries and Pushshift Reddit API (Baumgartner, 2019). These channels were chosen to match the terms in the second dataset. The dataset is divided into 2 different csv files, separating the posts made before the pandemic (January 2019 through April 2019) from those made during the pandemic (January 2020 through April 2020).

   Sample data (first 2 rows):

author	**created_utc**	**selftext**	**subreddit**	**title**
ant_collector5	1549402117	I feel like that guy from bench-warmers that's afraid of the sun.	depression	Haven't left bed in days

Computational Overview

1 Plot Line Graphs for Search Terms

The display_line_graphs.py module contains three functions. The read_xslx_file function takes in two strings, filename and term, as parameters. Using pandas and openpyxl packages, the function reads the xlsx file named as filename and extracts information listed in the column with term as the column name (Pandas User Guide, 2021). This column is returned as a dictionary, mapping the term string to the list containing all the values in the relevant column. The create_line_plots function takes in term (str) as the parameter variable. Using read_xslx_file as a helper function, the function then extracts data for a specific mental health-related term for all 6 countries and the worldwide data as provided in the search terms dataset. The function then creates and returns the line object generated by plotly.express (Line Charts in Python, 2021).
The plot_line_graphs function plots interactive line graphs using create_line_plots as a helper function. Using the custom buttons functionality of plotly, the user can choose which plot they want to see out of all 9 mental health-related search terms (Custom Buttons in Python, 2021). To accomplish this, we extracted individual traces from each line plot returned by create_line_plots and then stored them in a list named data. For each button, a list of boolean values were created, and each boolean value matched a trace in the data list, where True indicates that this trace line should be displayed when the button is clicked and vice versa for False. We then passed these boolean lists into plotly's update_layout method for every button so that each button works as intended.

2 Scrape Reddit Data and Process Texts

The reddit_scrape.py module contains just one function—scrape_subreddit_posts—dedicated to scraping posts from Reddit and writing all retrieved posts into one csv file using the pandas package (Pandas User Guide, 2021). This function takes in 4 parameters: after and before are timestamps that represent time in Coordinated Universal Time, subreddits is a tuple of strings, and filename is a string. With the help of the PushshiftAPI from the pmaw package, the function retrieves a maximum of 10,000 posts from each subreddit in the subreddits tuple that was created within the given timeframe, as defined by 'after' and 'before' timestamps (Baumgartner, 2019). The process_text.py module includes 3 functions, all dedicated to performing some processing on a given string (read the details in their respective docstrings). The PREDEFINED_CONTRACTIONS global variable stores commonly used contractions in English and their fully expanded form for the eliminate_contractions function to use (Enchanted Learning Contractions, 2018).

3 RedditPost and Subreddit Classes

The post.py module stores the RedditPost data class, the Subreddit class, a global variable STOPWORDS, and 2 top level functions related to posts. Representing a Reddit post, a RedditPost object has 5 instance attributes where the title and text of the post are stored as an array of strings that represent individual sentences. The Subreddit class represents an actual subreddit channel, so it has 2 private instance attributes (the channel name and a list of RedditPost objects).
The methods in the Subreddit class include words_frequency, intensity_analysis, and word_cloud. Using helpers from process_text.py, words_frequency takes in a set of keywords and returns a dictionary mapping each keyword to the number of times the keyword appears in the list of RedditPosts in the subreddit. With the help of SentimentIntensity-Analyzer from nltk.sentiment and the pre-trained VADER (Valence Aware Dictionary and sEntiment Reasoner) tool, intensity_analysis calculates the emotional intensity of each RedditPost in the subreddit (Bird et al., 2009). Lastly, word_cloud uses the generate method in WordCloud to create a word cloud image for the subreddit and stores it as a png file inside the "img" folder (Vu, 2019). Both intensity_analysis and word_cloud undergo text filtering using process_text.py helpers before feeding the filtered text into functions from the nltk and wordcloud packages. STOP-WORDS is a set of strings created by aggregating a standard list of stopwords compiled by Microsoft and a list of stopwords we created by observing output images to manually filter out common terms that are not already covered by the standard list (Rodionova, 2019). The two top-level functions perform polarity analysis (given a collection of intensity values) and compute average (given a list of values). These are not designed as methods of the class since they don't need self.

4 Load and Filter Data for Linear Regression

The load_regression_data.py module contains helper functions to load data from xlsx and csv files, filter them for the columns relevant to this project, and transform the extracted to data to be more easily-usable by storing them in

Python data structures. Two datasets are relevant for the linear regressions of this project.

- Data for COVID cases are stored as csv files, so we developed 3 helper functions (process_row, add_all_cases, and str_to_date) to help filter and transform the raw data into a computable form as a dictionary mapping datetime.date objects to ints that represent the aggregated total global confirmed cases on that day. Specifically, process_row eliminates columns with irrelevant geographical information, add_all_cases computes the sum of confirmed cases from every observation on one date (one column in the csv file), and str_to_date transforms the date in string "mm/dd/yy" format to datetime.date.
- Data for search terms are stored as xlsx files. We used the pandas package to read them and the iloc and loc methods to select the relevant rows; our function then returns a dictionary mapping datetime.date objects to a list that represents the number of searches for each mental health-related term in the data file for that day (Pandas User Guide, 2021).

filter_dataset function then filters these two dictionaries to make sure that the datetime.dates (keys) are identical for both datasets and returns them in a list. Data from Janurary 26, 2020 to May 31, 2020 were used.

5 Generate Linear Regressions

There are 3 global variables in the lin_regression.py module: FILTERED_SEARCH_INTEREST, FILTERED_GLOBAL_CASES, and MH_SEARCH_TERMS. The first two are dictionaries created by calling the filter_dataset helper function from load_regression_data.py, and the third is a list of strings representing all the mental health-related terms in the dataset for search terms. Our plot_graph function generates the linear regression for a mental health search term, which is the string parameter passed into the function. Using the numpy library, the x-axis and y-axis values are created as the number of global cases and search interest, respectively (NumPy Fundamentals, 2021). With scikit-learn functions, like fit and predict, a linear regression model is then produced for the number of confirmed cases worldwide versus global public attention towards a mental health related search term, which is quantified as Google search interest (Pedregosa et al., 2011). The matplotlib library and matplotlib.widgets helped us visualize the regression models and create a menu of buttons to organize the 9 graphs (one for each search term) in an interactive way (Matplotlib Widgets, 2021). Clicking a button opens a new window that shows the regression of the corresponding term; the regression coefficients, mean squared error, and coefficient of determination are also calculated and printed in the console. In this way, we could explore whether the number of confirmed Covid cases and the Google search interest for mental health-related terms are related in any statistically significant way (Pedregosa et al., 2011).

Instructions for Running the Program

1 Setting Up

Please download the appropriate Python packages as specified in `requirements.txt` and two datasets (hyperlinked): Global COVID-19 Cases and Global Mental Health Search Terms. Create 4 empty folders inside the main project folder and name them as "search_terms", "covid_data", "data", and "img". Store all files from the search terms dataset in search_terms and all files from the COVID-19 cases dataset in covid_data. Please ensure that all .py files are in the same main project folder.

(Advised) Download the Reddit dataset (hyperlinked), unzip it, and drag the 2 csv files inside into the data folder. We advise that you do this; though our code allows you to scrape Reddit data directly to recreate the same 2 csv files, it can take long (around 15 minutes), and sometimes the PushShiftAPI might actually reject some requests since we are scraping a large amount of post. If you do download this dataset directly from the link, please comment out the following section in main.py.

```
# Scraping posts data from the relevant subreddits
print("Scraping data from reddit...")
scrape(timestamps[0], timestamps[1], subreddits, 'before')
print("Completed for the first given timeframe...")
scrape(timestamps[2], timestamps[3], subreddits, 'after')
print("Reddit data has been acquired and saved as csv files inside the data folder")
```

2 Running main.py

Part 1: Line Graph

First, a page should open in your browser for you to see a graph similar to Figure 1. Use the dropdown on the left to see separate plots for specific search terms, such as for "depression" in Figure 2. Have fun exploring! You can use this interactive feature to see the variance in the search interest of different mental health terms across different countries before and during the pandemic.

Figure 1: Default Graph Containing Data On All 7 Countries For All 9 Terms

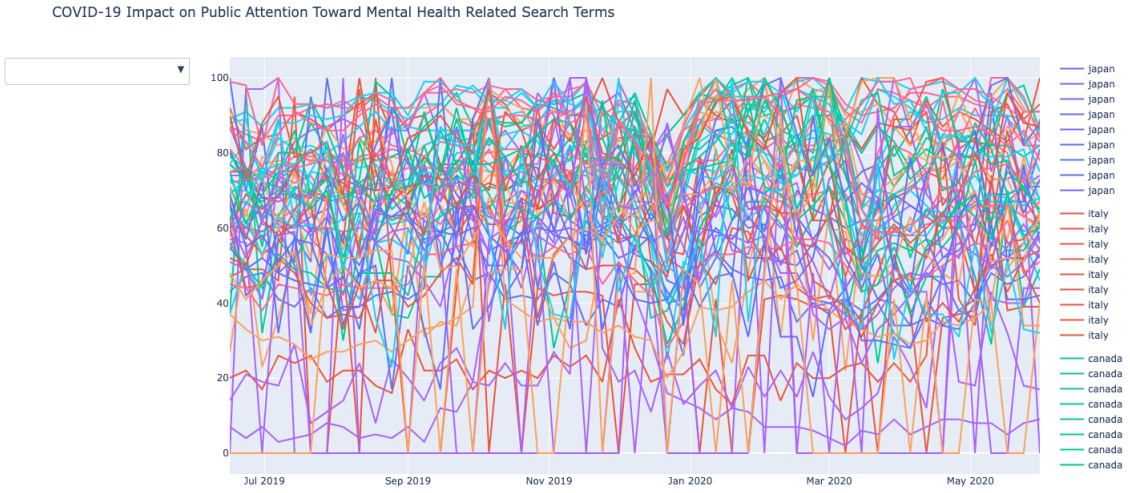
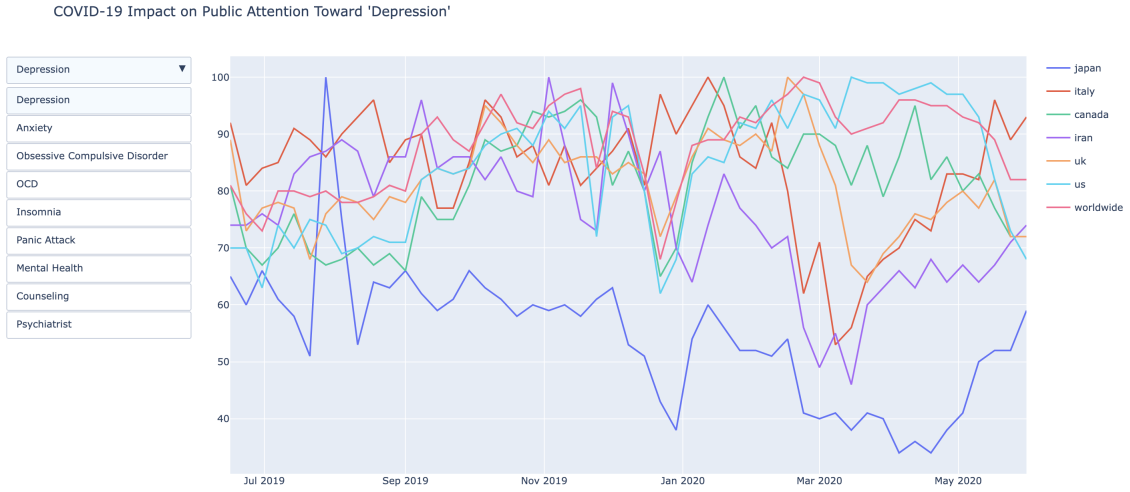


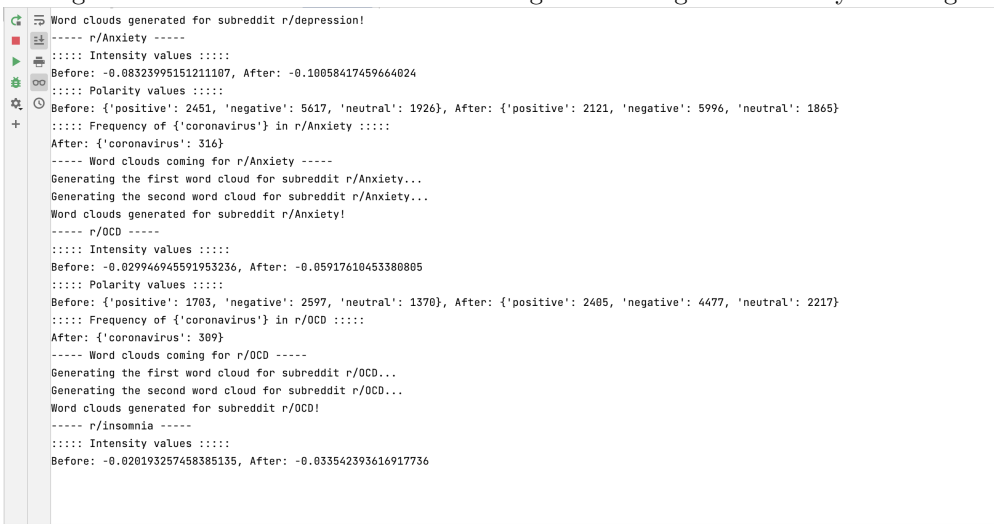
Figure 2: Filtered Graph that Contains Data On The Search Term 'Depression' For All 7 Countries



Part 2: Subreddit Sentiment Analysis and Word Clouds

Print statements will show up in the console to inform you which stage is currently running (Figure 3). The program starts with retrieving posts from all 8 subreddits for two different timeframes—1/1/2019 to 4/30/2019 and 1/1/2020 to 4/30/2020, or before and during the coronavirus outbreak. If you did not download the Reddit dataset directly, it should take around 15 minutes to scrape all the data from Reddit as we are processing tens of thousands of posts and writing them into local csv files (stored in the data folder). After successfully scraping, more print statements will show up in the console with information about the computed polarity and intensity values for each subreddit. Finally, the program then generates word clouds, which should be viewable in the img folder once the program finishes running. The word clouds for r/depression, r/Anxiety, and r/mentalhealth take relatively longer to generate (1-2 minutes) as these are the 3 most popular subreddits. Have fun exploring the word cloud image outputs inside the "img" folder and see if the sentiment analysis outputs match your expectations!

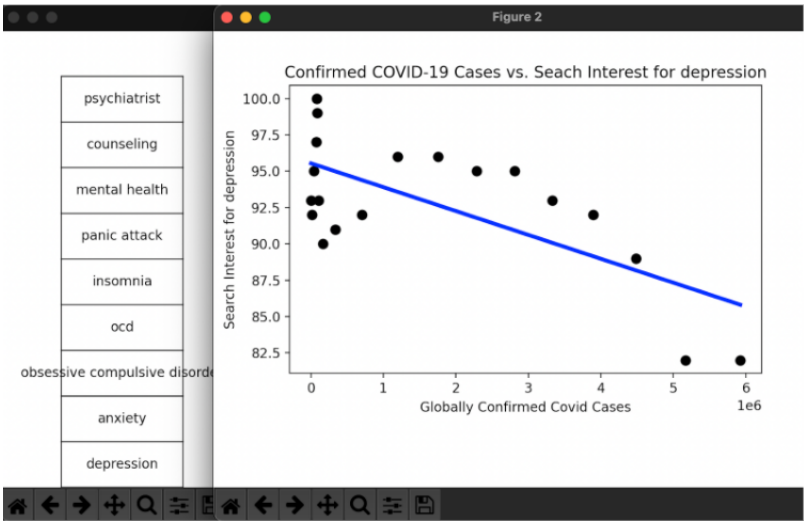
Figure 3: Print Statements Demonstrating Which Stage Is Currently Running



Part 3: Linear Regression Graph

Finally, two pop-up windows will open, one for the buttons (to switch between different regressions) and another for displaying the actual linear regressions (Figure 4). When a button is clicked, the regression coefficients, mean squared error, and coefficient of determination are also calculated and printed onto the console.

Figure 4: Search Term 'Depression' Linear Regression Graph Displayed Alongside All Search Term Buttons



Adjustments to the Original Proposal

In the proposal phase of the project, we planned to generate clusters plots for each subreddit to show the difference in the number of positive, negative, and neutral posts before and during the pandemic. We also wanted to create line graphs that plot intensity values of posts over time. However, we ended up creating word clouds for each subreddit to visualize the most frequently mentioned terms in the posts scraped from Reddit. We believe that word clouds is a more creative, interesting visualization than what we had previously planned.

Discussion and Conclusion

1 Results

1 Search Terms Line Graphs

There is one noticeable pattern across almost all search terms—the spike in search interest during the early weeks of the pandemic, from January up until February 2020. This may be surprising especially in countries in North America and Europe as the awareness of COVID-19 started growing around the same time in March. The high search interest rates during this time in these countries could be interpreted in terms of the influence of changing weather patterns. For example, winter-related sicknesses are high during January and February in North America, so it is expected that the search interest would be high. However, during the transition from winter to spring, the search interest would be expected to be lower as winter-related sicknesses begin to wane–except in this case, COVID-19 boosted the search interest despite the seasonal change in the March to May window. Italy clearly shows this trend. For example, search interest of "anxiety" remained high during the off-peak spring season from March to May as Italy had one of the highest death tolls from COVID-19. Similarly, UK and Canada had high search interest during this time but lower than Italy since they were not in as bad of a situation in terms of experiencing the effects of the pandemic. A similar trend can be observed for the term "insomnia". Moreover, for some terms and countries, the largest peaks occurred later on in the pandemic around March and even May, such as the spike for "OCD" in May.

Depression: Canada, US, Italy, and UK witnessed the greatest spikes in January and early February. Interestingly, search interest in Iran dipped significantly after the common spike in January. Worldwide, the highest peaks occurred in late February and early April.

Anxiety: Search interest spiked in January and February for all countries other than Iran, which witnessed the spike in March, April and May. Though interest increased in Japan throughout the pandemic, values remained low (relative to other countries).

Obsessive Compulsive Disorder: All countries witnessed increases, but the interest generally stayed consistent for Japan. Iran, UK, and Italy kept fluctuating between very high and very low values, making it hard to interpret.

OCD: spiked for all countries and worldwide in January and/or May.
Insomnia: Japan, UK, Canada, and US all had spikes in search interest in January/February. Also, unlike most other countries, Italy and Irans' search interest peaked in March and/or April.

Panic Attack: Aside from Japan, all countries and worldwide had spikes in March and/or April. Iran's search interest showed significant fluctuation between very high and low values compared to other countries.

Mental Health: Italy, Canada, UK, and the US all had peaks in May. Italy, Canada, Japan, and US all had peaks in January/February. Iran's search interest was considerably low throughout the pandemic and showed no significant peaks.

Counseling: All countries had spikes in January and February. Canada, UK, US and worldwide showed a very similar trend; they all had significant peaks in January, February, and March, followed by significant drops.

Psychiatrist: All countries and worldwide had peaks in search interest in January and/or February. Italy was the only country that had significant peaks in the months January, February and May; other countries had their most significant spikes only in January and/or February.

2 Sentiment Analysis and Word Clouds

Part 1: Reddit Sentiment Analysis
For intensity values, we computed the "compound" value yielded from the SentimentAnalyzer in nltk.sentiment. This value is an aggregated metric calculated by normalizing the sum of all intensity values (for positive, negative, and neutral sentiments) between -1 (most negative) and +1 (most positive). We can conclude that except r/counselling, r/mentalhealth, and r/PanicAttack, there is a slight decrease in the compound scores for all remaining channels, meaning that the overall sentiment was more negative during the pandemic than before. However, when looking at dictionary outputs that map polarity values to the total number of posts, we can draw a few more conclusions. As shown in Table 1:

- r/depression and r/Anxiety are the only two subreddits where the number of negative posts have went up while the number of positive and neutral posts decreased.
- There has been an increase in all types of posts in r/OCD, r/insomnia, r/PanicAttack, and r/Psychiatry, potentially pointing toward the fact that there was more activity overall on social media during the pandemic.
- Interestingly, r/mentalhealth is the only subreddit where the number of negative posts have gone down while the number of positive posts increased.
- All types of posts decreased in r/counselling.

Table 1: Results of Valence-based and Polarity-based Sentiment Analysis and Frequency of Covid-related Terms

r/	**before-intensity**	**after-intensity**	**before-polarity**	**after-polarity**	**frequency of the term 'coronavirus'**
depression	-9.269×10^{-2}	-1.009×10^{-1}	'positive': 2370, 'negative': 5662, 'neutral': 1933	'positive': 2361, 'negative': 5766, 'neutral': 1836	101
Anxiety	-8.324×10^{-2}	-1.006×10^{-1}	'positive': 2451, 'negative': 5617, 'neutral': 1926	'positive': 2121, 'negative': 5996, 'neutral': 1865	316
OCD	-2.995×10^{-2}	-5.918×10^{-2}	'positive': 1703, 'negative': 2597, 'neutral': 1370	'positive': 2405, 'negative': 4477, 'neutral': 2217	309
insomnia	-2.019×10^{-2}	-3.354×10^{-2}	'positive': 532, 'negative': 704, 'neutral': 413	'positive': 809, 'negative': 1296, 'neutral': 739	18
PanicAttack	-2.300×10^{-1}	-2.140×10^{-1}	'positive': 56, 'negative': 526, 'neutral': 71	'positive': 88, 'negative': 608, 'neutral': 92	20
mentalhealth	-6.370×10^{-2}	-6.367×10^{-2}	'positive': 2607, 'negative': 5240, 'neutral': 2149	'positive': 2768, 'negative': 5195, 'neutral': 2029	160
counselling	7.790×10^{-2}	8.329×10^{-2}	'positive': 58, 'negative': 28, 'neutral': 47	'positive': 46, 'negative': 16, 'neutral': 34	5
Psychiatry	2.090×10^{-2}	2.340×10^{-3}	'positive': 257, 'negative': 181, 'neutral': 168	'positive': 271, 'negative': 225, 'neutral': 234	6

A table containing the before-intensity, after-intensity, before-polarity, after-polarity, and frequency of the term 'coronavirus' values. The term 'before' refers to the the first timeframe (January 2019 through April 2019), while the term 'after' refers to the second timeframe (Janauary 2020 through April 2020)

Part 2: Word cloud observations
Comparing the output images to the frequency values shown in Table 1, we can see that 'coronavirus' has definitely been discussed in these subreddits, but the proportion of its frequency to the total number of posts is within roughly 5% across all 8 subreddits. This matches what we observe from looking at the word clouds generated for the

subreddits. For instance, we can see a small "coronavirus" term in Figure 5. The term is pretty faint and located directly below "reason", which is right below "friend". We cannot seem to see a similar occurrence for the other 7 subreddits, potentially due to the fact that there are many more posts in these subreddits than r/counselling.

Figure 5: Word Cloud for r/counselling (analyzing posts from 1/1/2020 to 4/30/2020)

3 Linear Regression for Confirmed Covid Cases and Search Terms

The prediction accuracy and correlation for the linear regression models vary significantly between each search term and do not show any particular pattern overall—models have either weak positive or weak negative correlations. A commonality across most plots is the highest search interest being in the initial few months where confirmed cases were low. This can be a surprising result since one would expect mental health to worsen with more time spent in quarantine and news of increasing confirmed cases.

However, as shown in Table 2, the correlations from the linear regression models are weak, evident in the low coefficients of determination, which implies that the regression model fails to explain the variability in search interest, and relatively high mean squared errors, which represents low prediction accuracy. Therefore, it can be concluded that the number of global confirmed cases does not significantly impact public attention on mental health (quantified here as search interest of mental-health related terms on Google Trends).

Table 2: Search Term Popularity on Google Trends vs. Confirmed Covid Cases Linear Regression Results

Search Term	Slope Coefficient	Mean Squared Error	Coefficient of Determination
depression	-1.64×10^{-6}	11.157	0.47
anxiety	-6.31×10^{-7}	9.089	0.14
obsessive compulsive disorder	-1.61×10^{-6}	76.417	0.11
ocd	5.12×10^{-6}	50.988	0.66
insomnia	2.38×10^{-6}	61.303	0.26
panic attack	-8.33×10^{-7}	40.822	0.06
mental health	2.67×10^{-7}	22.836	0.01
counseling	-3.86×10^{-6}	143.538	0.28
psychiatrist	-1.79×10^{-6}	165.167	0.07

A table containing the slope coefficient, mean squared error, and coefficient of determination of the regression for each search term.

2 Challenges and Limitations

Search Terms Line Graphs

Implementing interactive line plots was quite challenging at first as most of the resources on the web surrounded bar graphs and scatterplots. Thus, we had to read more plotly documentation to better understand the update_layout

method and custom buttons' functionality. The turning point came when we realized we had to extract individual line traces into a list for each search term, and as a result, each button had a large list of boolean values to fully inform plotly about which line(s) should be visible depending on which button is currently active.

Sentiment Analysis and Word Clouds for Subreddits
We introduced a maximum of 10,000 posts to scrape for each subreddit to avoid spending too much time on retrieving data as there is a limited number of request that the PushshiftAPI can make at once (Baumgartner, 2019). 10,000 posts should sufficiently represent the overall sentiment of each subreddit, for all subreddits (other than r/depression with 50,000 posts) had around or less than 10,000 posts. VADER handles data from social media platforms well as it considers capitalization, emojis, and emotional aspects of punctuation (e.g. exclamation marks), but it works better when analyzing one sentence at a time (Bird et al., 2009). When we realized that many users post long paragraphs, we used sent_tokenize from nltk.tokenize to separate long posts into individual sentences before passing them into our SentimentIntensityAnalyzer object one by one for analysis (Bird et al., 2009). The resulting intensity value of each post was calculated as the average of the intensity values of the sentences in the post. However, for other text analyses, such as calculating the frequency of a word and generating word clouds, we eventually realized that we need to do more careful processing of the text, such as turning all characters lowercase, expanding contractions, and eliminating non-alphabetic characters before calling on functions to analyze a piece of text.

Linear Regression for Confirmed Covid Cases and Search Terms
Although the first and second data sets are large independently, the usable parts of them are not. To fit a linear regression model, the dates of the confirmed cases need to align with the dates corresponding to when search interest was measured. However, the number of dates where the two datasets overlap amount to only 19 data points from January 26 to May 31. Thus, without a large amount of overlapping data, the reliability and accuracy of the regression model is weakened. Google is not a major search engine in some countries contained in data set 1 that largely contribute to the number of confirmed COVID-19 cases (e.g. China, South Korea). Therefore, the search interests at these locations are excluded, which may weaken the accuracy and limit the applicability of the model.

3 Next Steps

Search Terms Line Graphs
To be able to better interpret the impact of COVID-19 on public attention toward mental health, it would be useful to have data for all of 2019 and 2020 as opposed to only specific months of each year as that way we could see the cycle of the entire year for both years and thus have a more accurate comparison. This would allow us to exclude the influence of the changing weather patterns. For example, as seen in the line graphs, the most noticeable trend was the increase in search interest during the early weeks of the pandemic in January and February. This could reflect the fact that winter-related sicknesses during these months are high in North America for example rather than it being due to the pandemic since the effects of the pandemic in North America and in Europe were witnessed around the same time in March rather than early on in January and February. Also, our data went up until May of 2020, though the effects of the pandemic continued to be witnessed well past that time, hence the data for the full year would allow us to answer our research question with more clarity. Since we did not have the full data for 2019, we hypothesized that the search terms spike in January and February is winter-related, but if we had data for the full year of 2019, we would be able to conclude with more clarity that the peak in January and February is winter-related.

Sentiment Analysis and Word Clouds for Subreddits
The STOPWORDS lists in post.py can be updated with more words. This list is used to filter out common words from wordclouds, but it doesn't include modern slang like "lol" which can appear often in posts on Reddit. Another potential improvement here is detecting contraction typos. Our program currently handles filtering of contractions intelligently, but many Reddit posts contain terms like "im" and "youre" without the actual contraction, making it hard for our program to detect currently. We can also analyze subreddit comments beyond the posts by adding another function to the reddit_scrape.py module for scraping comments data. In this way, we might collect data from more Reddit users since some users prefer commenting rather than actually posting in subreddits.

Linear Regression for Confirmed Covid Cases and Search Termsn
In addition to analysing the impact of confirmed cases on public attention on mental health, death cases and recovered cases could also be used to fit regression models that won't be limited to being linear. Some plots resemble parabolic and other more complex curves, which would be better fit by non-linear regression models. Moreover, search interest of the terms would be collected from other major search engines like Baidu, Naver, and Bing so that search interest

can be applied to a global scale. The search interests would be averaged to find the global average search interest in mental health-related terms, which would likely produce a more accurate regression model with the number of global confirmed cases.

References

Andrews, T. M. (2020, March 24). Our IPhone Weekly Screen Time Reports Are through the Roof, and People Are 'Horrified'. *The Washington Post*. Retrieved from https://www.washingtonpost.com.

Baumgartner, J. M. (2019, October 01). Pushshift API. Github repository. Retrieved from https://github.com/.

Bird, S., Loper, E., & Klein, E. (2009). Natural Language Processing with Python. O'Reilly Media Inc.

COVID-19 Mental Disorders Collaborators. (2021, October 08). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *The Lancet*. Retrieved from https://www.thelancet.com/journals/lancet/home.

Custom Buttons in Python. (2021). plotly Graphing Libraries. Retrieved from https://plot.ly.

Dong, E., Du, H., & Gardner, L. (2020, February 19). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet*. Retrieved from https://www.thelancet.com/journals/lancet/home.

Enchanted Learning Contractions. (2018). EnchantedLearning.com. Retrieved from https://www.enchantedlearning.com/grammar/.

Government of Canada. (2021, October 28). Coronavirus disease (COVID-19): Prevention and risks. Retrieved from https://www.canada.ca/en.html.

Hao, Y. (2020). COVID-19 and Mental Health Search Terms, 16. Retrieved from https://www.kaggle.com/.

Kamal, R. K., Panchal, N., Cox, C., & Garfield, R. (2021, February 10). The Implications of COVID-19 for Mental Health and Substance Use. Kaiser Family Foundation. Retrieved from https://www.kff.org/.

Line Charts in Python. (2021). plotly Graphing Libraries. Retrieved from https://plot.ly.

Matplotlib Widgets. (2021). Matplotlib. Retrieved from https://matplotlib.org/.

Mohammad, S. M. (2016, April 15). Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. *ScienceDirect*. Retrieved from https://www.sciencedirect.com/.

NumPy Fundamentals. (2021). NumPy. Retrieved from https://numpy.org/.

Pandas User Guide. (2021). Pandas Documentation. Retrieved from https://pandas.pydata.org/.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12,* 2825-2830.

Plotly Technologies Inc. (2015). Collaborative data science. Retrieved from https://plot.ly.

Rodionova, E. (2019, October 16). PowerBI-visuals-WordCloud. Github repository under Microsoft. Retrieved from https://github.com/microsoft/.

Vu, D. (2019, November 8). Generating WordClouds in Python. Datacamp. Retrieved from https://www.datacamp.com.

World Health Organization. (2020, October 05). COVID-19 disrupting mental health services in most countries, WHO survey. Retrieved from https://www.who.int/.